

Multivariate Statistics in Ecology and Quantitative Genetics

Redundancy analysis

Dirk Metzler & Noémie Becker

http://evol.bio.lmu.de/_statgen

Summer semester 2013

1 Redundancy analysis

- Setting
- Example: Artificial fish data
- Triplots
- Example: Height weight data
- Example: Species richness on sandy beaches (RIKZ data)
- The order of importance

Contents

- 1 Redundancy analysis
 - Setting
 - Example: Artificial fish data
 - Triplots
 - Example: Height weight data
 - Example: Species richness on sandy beaches (RIKZ data)
 - The order of importance

Given: Data frames/matrices Y and X

The variables in X are called explanatory variables

The variables in Y are called response variables

- Given:** Data frames/matrices Y and X
The variables in X are called explanatory variables
The variables in Y are called response variables
- Goal:** Find those components of Y which are linear combinations of X and (among those) represent as much variance of Y as possible.

Given: Data frames/matrices Y and X

The variables in X are called explanatory variables

The variables in Y are called response variables

Goal: Find those components of Y which are linear combinations of X and (among those) represent as much variance of Y as possible.

Assumption: There is a linear dependence of the response variables in Y on the explanatory variables in X .

- Given:** Data frames/matrices Y and X
The variables in X are called explanatory variables
The variables in Y are called response variables
- Goal:** Find those components of Y which are linear combinations of X and (among those) represent as much variance of Y as possible.
- Assumption:** There is a linear dependence of the response variables in Y on the explanatory variables in X .

The idea behind redundancy analysis is to apply linear regression in order to represent Y as linear function of X and then to use PCA in order to visualize the result.

- Given:** Data frames/matrices Y and X
The variables in X are called explanatory variables
The variables in Y are called response variables
- Goal:** Find those components of Y which are linear combinations of X and (among those) represent as much variance of Y as possible.
- Assumption:** There is a linear dependence of the response variables in Y on the explanatory variables in X .

The idea behind redundancy analysis is to apply linear regression in order to represent Y as linear function of X and then to use PCA in order to visualize the result.

Among those components of Y which can be linearly explained with X (multivariate linear regression) take those components which represent most of the variance.

Before applying RDA:


- Is Y increasing with increasing values of X ?
- If the variables in X are twice as high, are the variables in Y also approximately twice as high?

These questions are to check the assumption of linear dependence.

Contents

- 1 Redundancy analysis
 - Setting
 - **Example: Artificial fish data**
 - Triplots
 - Example: Height weight data
 - Example: Species richness on sandy beaches (RIKZ data)
 - The order of importance

To illustrate the output of redundancy analysis (RDA) we consider an artificial example from p. 590 of

 P. Legendre and L. Legendre.
Numerical Ecology

(We will not go into the maths of RDA)

The artificial data set represents fish abundances at 10 sites along a tropical reef transect. The first three sites are on “sand” and the others alternate between “coral” and “other substrate”. The water depth is given in meter.

The artificial data set represents fish abundances at 10 sites along a tropical reef transect. The first three sites are on “sand” and the others alternate between “coral” and “other substrate”. The water depth is given in meter.

```
> fishes <- read.table("artificialFishes.txt",h=T); fishes
```

	Site	Sp1	Sp2	Sp3	Sp4	Sp5	Sp6	Depth	Coral	Sand	Other
1	1	1	0	0	0	0	0	1	0	1	0
2	2	0	0	0	0	0	0	2	0	1	0
3	3	0	1	0	0	0	0	3	0	1	0
4	4	11	4	0	0	8	1	4	0	0	1
5	5	11	5	17	7	0	0	5	1	0	0
6	6	9	6	0	0	6	2	6	0	0	1
7	7	9	7	13	10	0	0	7	1	0	0
8	8	7	8	0	0	4	3	8	0	0	1
9	9	7	9	10	13	0	0	9	1	0	0
10	10	5	10	0	0	2	4	10	0	0	1

The abundancies of the six species are the response variables.
'Depth', 'Coral', 'Sand' and 'Other' are explanatory variables.
We do not need to `scale=TRUE` as abundancies are on comparable scales.

The abundancies of the six species are the response variables. 'Depth', 'Coral', 'Sand' and 'Other' are explanatory variables. We do not need to `scale=TRUE` as abundancies are on comparable scales. As 'Coral', 'Sand' and 'Other' are linearly dependent, the covariance matrix is singular. So we can only use two out of the three variables.

The abundancies of the six species are the response variables. 'Depth', 'Coral', 'Sand' and 'Other' are explanatory variables.

We do not need to `scale=TRUE` as abundancies are on comparable scales.

As 'Coral', 'Sand' and 'Other' are linearly dependent, the covariance matrix is singular. So we can only use two out of the three variables.

We choose 'Depth', 'Sand' and 'Coral' as explanatory variables.

The abundancies of the six species are the response variables. 'Depth', 'Coral', 'Sand' and 'Other' are explanatory variables.

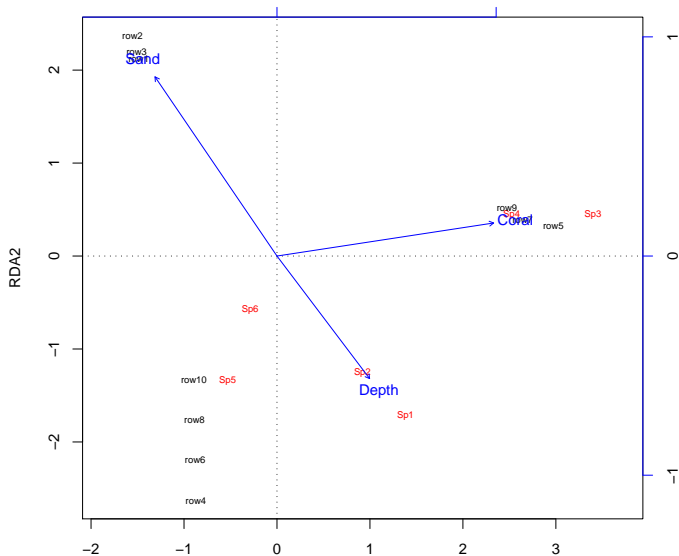
We do not need to `scale=TRUE` as abundancies are on comparable scales.

As 'Coral', 'Sand' and 'Other' are linearly dependent, the covariance matrix is singular. So we can only use two out of the three variables.

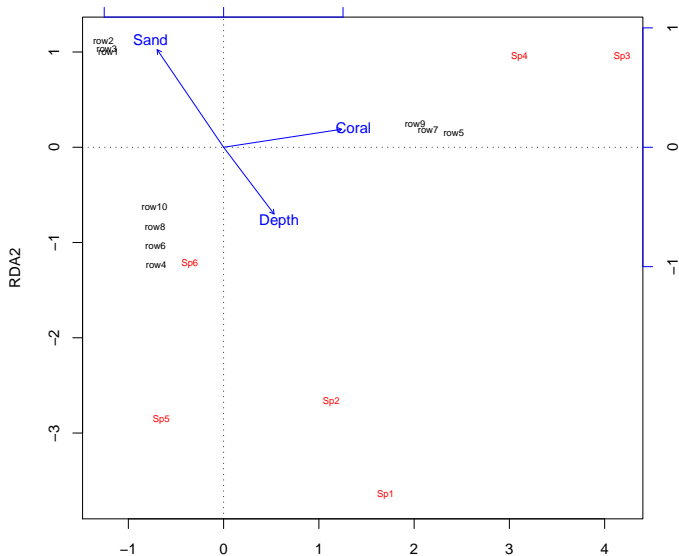
We choose 'Depth', 'Sand' and 'Coral' as explanatory variables.

```
library(vegan) # rda() is in this library
Resp <- fishes[,c("Sp1", "Sp2", "Sp3", "Sp4", "Sp5", "Sp6")]
Expl <- fishes[,c("Depth", "Coral", "Sand", "Other")]
myrda <- rda(Resp, Expl)
plot(myrda, scaling=2)
plot(myrda, scaling=1)
```

Correlation triplot



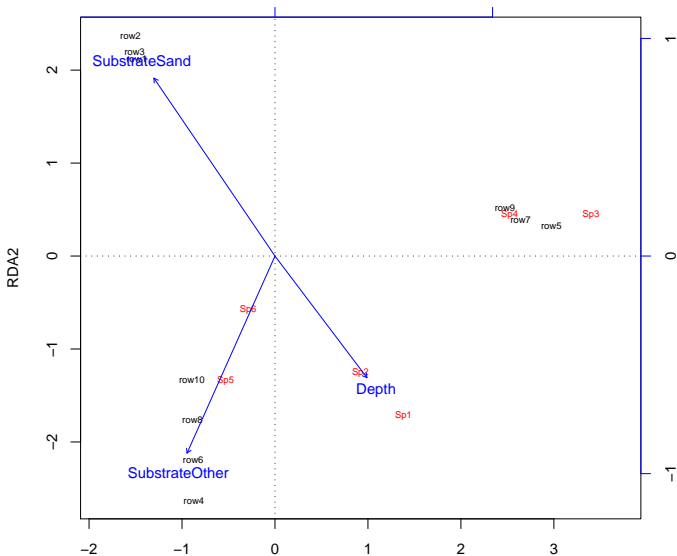
Distance triplot



If we gather the three variables 'Coral', 'Sand' and 'Other' into one factor variable 'Substrate', then R eliminates automatically the last variable.

```
Substrate <- c(rep("Sand",3),  
              rep(c("Other","Coral"),3),"Other")  
myrda <- rda(Resp~Depth+Substrate,data=Expl)  
plot(myrda,scaling=2)
```

Correlation triplot:



Contents

- 1 Redundancy analysis
 - Setting
 - Example: Artificial fish data
 - **Triplots**
 - Example: Height weight data
 - Example: Species richness on sandy beaches (RIKZ data)
 - The order of importance

The graphical output of RDA consists of two biplots on top of each other and is called `triplot`.

You produce a triplot with `plot(rda.object)` (which itself calls `plot.cca()`).

The graphical output of RDA consists of two biplots on top of each other and is called `triplot`.

You produce a triplot with `plot(rda.object)` (which itself calls `plot.cca()`).

There are three components in a triplot:

- Continuous explanatory variables (numeric values) are represented by lines. Nominal explanatory variables (factor object) (coded 0 – 1) by squares (or triangles) (one for each level). The square is plotted at the centroid of the observations that have the value 1.

The graphical output of RDA consists of two biplots on top of each other and is called `triplot`.

You produce a triplot with `plot(rda.object)` (which itself calls `plot.cca()`).

There are three components in a triplot:

- Continuous explanatory variables (numeric values) are represented by lines. Nominal explanatory variables (factor object) (coded 0 – 1) by squares (or triangles) (one for each level). The square is plotted at the centroid of the observations that have the value 1.
- The response variables by labels or lines.

The graphical output of RDA consists of two biplots on top of each other and is called `triplot`.

You produce a triplot with `plot(rda.object)` (which itself calls `plot.cca()`).

There are three components in a triplot:

- Continuous explanatory variables (numeric values) are represented by lines. Nominal explanatory variables (factor object) (coded 0 – 1) by squares (or triangles) (one for each level). The square is plotted at the centroid of the observations that have the value 1.
- The response variables by labels or lines.
- The observations by points or labels.

Distance triplot (scaling=1)

- Distances between points (observations), between squares or between points and squares approximate distances of the observations (or the centroid of the nominal explanatory variable).
- Angles between lines of response variables and lines of explanatory variables represent a two-dimensional approximation of correlations.
- Other angles between lines are meaningless.

Distance triplot (scaling=1)

- Distances between points (observations), between squares or between points and squares approximate distances of the observations (or the centroid of the nominal explanatory variable).
- Angles between lines of response variables and lines of explanatory variables represent a two-dimensional approximation of correlations.
- Other angles between lines are meaningless.
- The projection of a point onto the line of a response variable at right angle approximates the position of the corresponding object along the corresponding variable.
- Squares/triangles cannot be compared with lines of qualitatively explanatory variables.

Correlation triplot (scaling=2)

- The cosine of the angle between lines (of response variable or of explanatory variable) is approximately equal to the correlation between the corresponding variables.

Correlation triplot (scaling=2)

- The cosine of the angle between lines (of response variable or of explanatory variable) is approximately equal to the correlation between the corresponding variables.
- Distances are meaningless.

Correlation triplot (scaling=2)

- The cosine of the angle between lines (of response variable or of explanatory variable) is approximately equal to the correlation between the corresponding variables.
- Distances are meaningless.
- The projection of a point onto a line (response variable or explanatory variable) at right angle approximates the value of the corresponding variable of this observation.

Correlation triplot (scaling=2)

- The cosine of the angle between lines (of response variable or of explanatory variable) is approximately equal to the correlation between the corresponding variables.
- Distances are meaningless.
- The projection of a point onto a line (response variable or explanatory variable) at right angle approximates the value of the corresponding variable of this observation.
- The length of lines are not important.

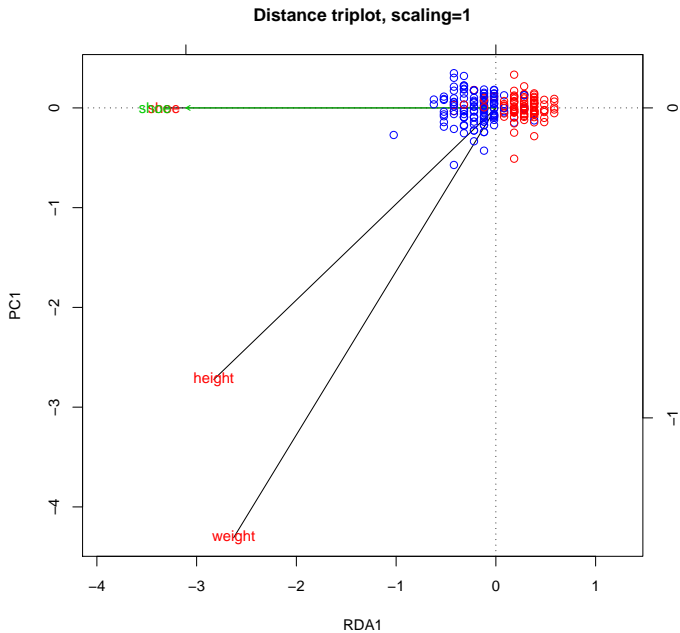
Contents

- 1 Redundancy analysis
 - Setting
 - Example: Artificial fish data
 - Triplots
 - **Example: Height weight data**
 - Example: Species richness on sandy beaches (RIKZ data)
 - The order of importance

Recall `hsw` and `fm.col` from the slides on PCA.

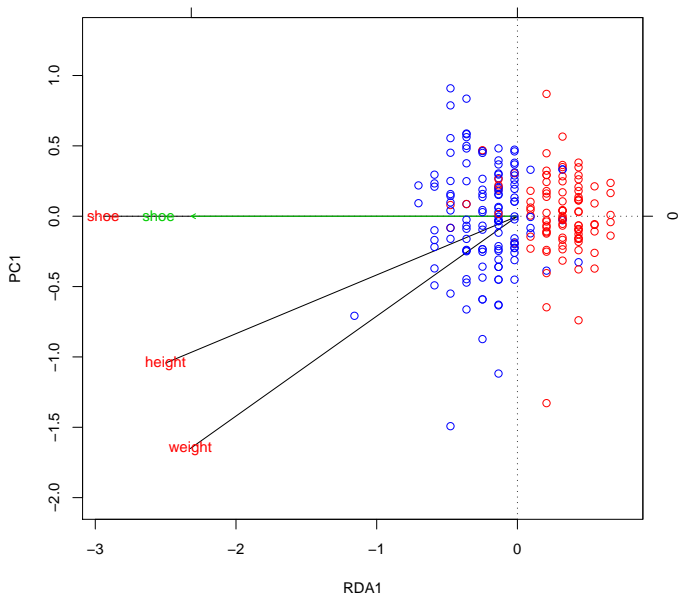
```
Expl <- hsw[,c(1,2,3)]
Resp <- hsw[,c(1,2,3)]
myrda <- rda(Resp~shoe,scale=TRUE,data=Expl)

# Distance triplot
# The following command does not plot (type=None)
plot(myrda,scaling=1,type="n",main="Distance triplot")
segments(x0=0,y0=0,
         x1=scores(myrda, display="species", scaling=1)[,1],
         y1=scores(myrda, display="species", scaling=1)[,2])
text(myrda, display="sp", scaling=1, col=2)
text(myrda, display="bp", scaling=1,
     row.names(scores(myrda, display="bp")), col=3)
points(myrda,display=c("sites"),scaling=1,pch=1,col=fm.col)
```



```
# Correlation triplot
plot(myrda,scaling=2,type="n",main="Correlation triplot")
segments(x0=0,y0=0,
         x1=scores(myrda, display="species", scaling=2)[,1],
         y1=scores(myrda, display="species", scaling=2)[,2])
text(myrda, display="sp", scaling=2, col=2)
text(myrda, display="bp", scaling=2,
     row.names(scores(myrda, display="bp")), col=3)
points(myrda,display=c("sites"),scaling=2,pch=1,col=fm.co
```

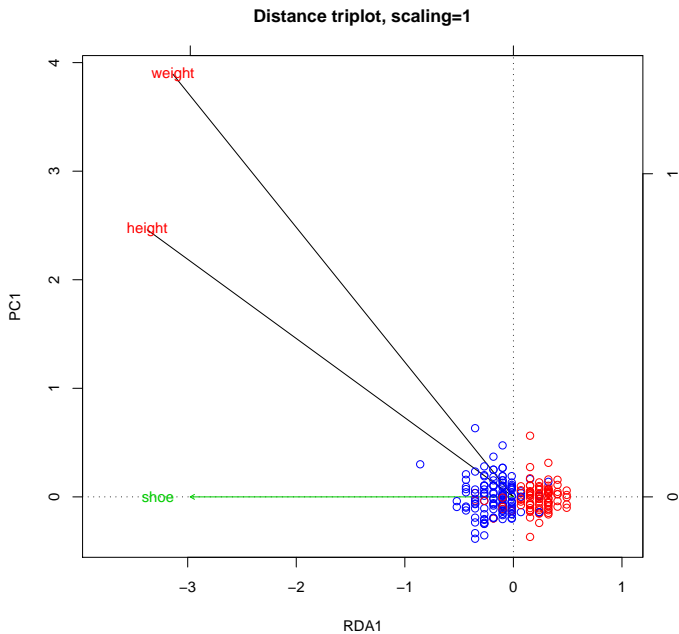
Correlation triplot, scaling=2



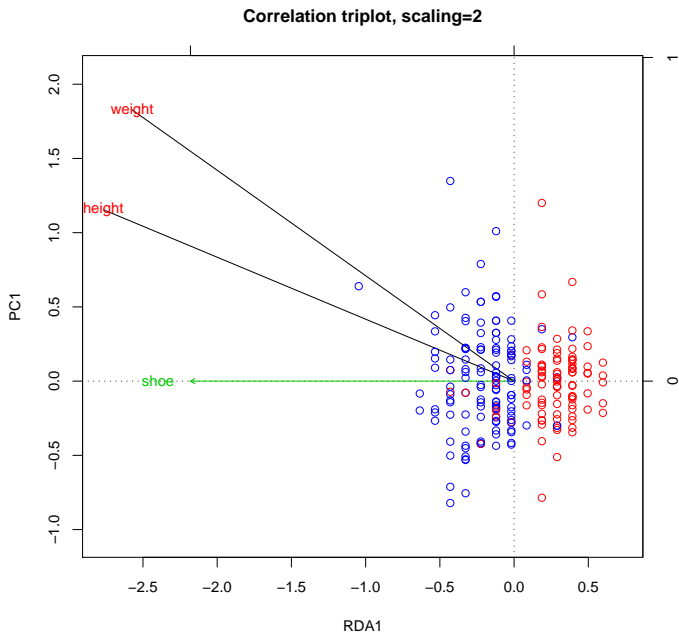
Now without shoe as response variable:

```
Expl <- hsw[,c(1,2,3)]
Resp <- hsw[,c(1,3)]
myrda <- rda(Resp~shoe,scale=TRUE,data=Expl)

# Distance triplot
# The following command does not plot (type=None)
plot(myrda,scaling=1,type="n",main="Distance triplot")
segments(x0=0,y0=0,
         x1=scores(myrda, display="species", scaling=1)[,1],
         y1=scores(myrda, display="species", scaling=1)[,2])
text(myrda, display="sp", scaling=1, col=2)
text(myrda, display="bp", scaling=1,
     row.names(scores(myrda, display="bp")), col=3)
points(myrda,display=c("sites"),scaling=1,pch=1,col=fm.co
```



```
# Correlation triplot
plot(myrda,scaling=2,type="n",main="Correlation triplot")
segments(x0=0,y0=0,
         x1=scores(myrda, display="species", scaling=2)[,1],
         y1=scores(myrda, display="species", scaling=2)[,2])
text(myrda, display="sp", scaling=2, col=2)
text(myrda, display="bp", scaling=2,
     row.names(scores(myrda, display="bp")), col=3)
points(myrda,display=c("sites"),scaling=2,pch=1,col=fm.co
```

Contents

- 1 Redundancy analysis
 - Setting
 - Example: Artificial fish data
 - Triplots
 - Example: Height weight data
 - **Example: Species richness on sandy beaches (RIKZ data)**
 - The order of importance

- Which factors influence the species richness on sandy beaches?
- Data from the dutch National Institute for Coastal and Marine Management (RIKZ: Rijksinstituut voor Kust en Zee)
- see also
 - 📖 Zuur, Ieno, Smith (2007) *Analysing Ecological Data*. Springer

	richness	angle2	NAP	grainsize	humus	week
1	11	96	0.045	222.5	0.05	1
2	10	96	-1.036	200.0	0.30	1
3	13	96	-1.336	194.5	0.10	1
4	11	96	0.616	221.0	0.15	1
.
.
21	3	21	1.117	251.5	0.00	4
22	22	21	-0.503	265.0	0.00	4
23	6	21	0.729	275.5	0.10	4
.
.
43	3	96	-0.002	223.0	0.00	3
44	0	96	2.255	186.0	0.05	3
45	2	96	0.865	189.5	0.00	3

Meaning of the Variables

index i index of sampling station

richness Number of species that were found in a plot.

angle1 angle of the station

angle2 slope of the beach at the plot

exposure index composed of wave action etc.

NAP altitude of the plot compared to the mean sea level.

grainsize average diameter of sand grains

humus fraction of organic material

week in which of 4 weeks was this plot probed.

(many more variables in original data set)

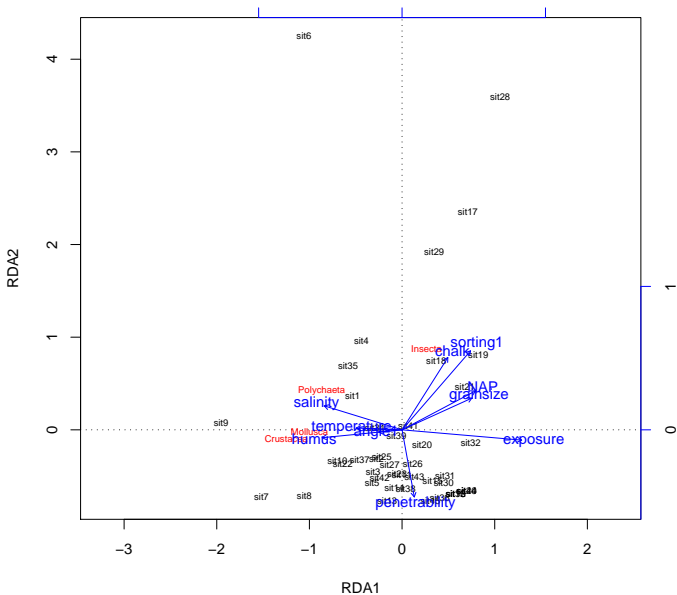
```
library(vegan)
RIKZ <- read.table("RIKZGroups.txt", header = TRUE)
Species <- RIKZ[,2:5]
#Data were square root transformed
Species.sq <- sqrt(Species)

I1 <- rowSums(Species) #Could be used to drop sites with
                        #of 0.

ExplVar <- RIKZ[, c("angle1", "exposure", "salinity",
                   "temperature", "NAP", "penetrability",
                   "grainsize", "humus", "chalk",
                   "sorting1")]

RIKZ_RDA <- rda(Species.sq, ExplVar, scale=T)
plot(RIKZ_RDA, scaling=2)
```

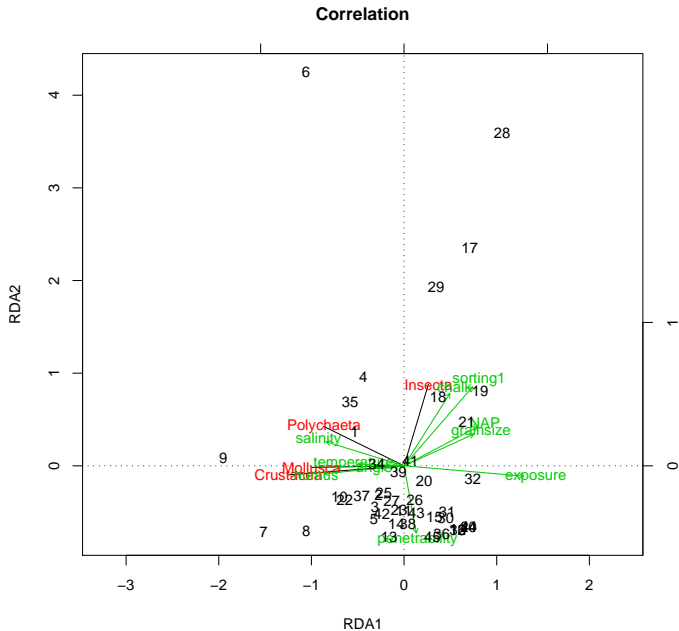
Correlation biplot



A different triplot

```
# Correlation biplot, scaling=2
plot(RIKZ_RDA, scaling=2, main="Correlation", type="n")
segments(x0=0, y0=0,
         x1=scores(RIKZ_RDA, display="species", scaling=2),
         y1=scores(RIKZ_RDA, display="species", scaling=2))
text(RIKZ_RDA, display="sp", scaling=2, col=2)
text(RIKZ_RDA, display="bp", scaling=2,
     row.names(scores(RIKZ_RDA, display="bp")), col=3)
text(RIKZ_RDA, display=c("sites"), scaling=2, labels=row.names(scores(RIKZ_RDA, display="sites")))

cor(Species.sq, ExplVar)
```

Contents

- 1 Redundancy analysis
 - Setting
 - Example: Artificial fish data
 - Triplots
 - Example: Height weight data
 - Example: Species richness on sandy beaches (RIKZ data)
 - **The order of importance**

Anova on RDA objects

Which of the explanatory variables is the most important?

Which are the least important or even irrelevant?

As RDA is based on linear regression, the same methods apply.

Due to time constraint, this is not part of the lecture.

Anova on RDA objects

Which of the explanatory variables is the most important?
Which are the least important or even irrelevant?

As RDA is based on linear regression, the same methods apply.
Due to time constraint, this is not part of the lecture.

Have a try for yourself:

```
anova(RIKZ_RDA)  
step(RIKZ_RDA)  
dropterm(RIKZ_RDA)
```